# Internship Assignment

The assignment was given with 2 different tasks:

1. Classification of industries according to:
   * + - 1. StartUps: Started after 1990; tech-oriented and innovative
         2. University/Schools
         3. Govt./Non-Profit
         4. Unclassified
         5. Mature: Started before 1990
2. Dataset given was clean, however it was needed to be cleaned further, since:
   * + - 1. Launch Date had Month(for some rows) and Year. However, since we only required Year of Launch, so I split the column into Launch Year using PowerQuery and removed Launch Month
         2. TAGS and TAGLINE were two important parameters required to make classification. However, data was having there were missing TAGS and TAGLINES for some companies and data was split into 3:

With only TAGS

With only TAGLINES

Both TAGS and TAGLINES

1. TAGS were split into list of tags . And separate list were created for Startups, government and education related words. TAGS were compared with these list based on SET interaction. If list of tags and classified lists had mutual words, then they were classified accordingly.
2. Similar exercise was done with TAGLINE. However, to create list of words in taglines, taglines was broken down into tokens and list was created after removing stop words.
3. Count of Industries in datasets are as follows:

TYPE

Startup 7174

Unclassified 2311

Government/Non-Profit 1514

Universities/Schools 450

Mature 88

**Assignment 2: Web scrapping : Ycombinator.com/companies**

WebScrapping is done using Selenium. I used Xpaths for getting attributes of every class. This data frame contains 3 Columns: Company Name, Company Information, and Website link